

Giant extrachromosomal element “Inocle” potentially expands the adaptive capacity of the human oral microbiome

Corresponding Author: Professor Yutaka Suzuki

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

In this study, the authors report the discovery of Inocles, a novel intracellular plasmid-like extrachromosomal element of *Streptococcus*. These novel features have several interesting properties and appear to have correlations with certain immune cells and proteins, and negative correlations with several cancers. Aside from some sporadic grammatical errors, the report is well-written, the approach is logical and well-conducted, the conclusions are largely supported by the results, and my enthusiasm for this article is high. There are several shortcomings of the current report, and addressing these would strengthen the document. Specific points to address are the following:

1. It is not entirely clear to this reviewer how Inocles differ from plasmids (other than that recent improvements in long-read sequencing technology enabled their discovery). Clearly Inocles are novel, but it would be helpful for the reader to have a more straightforward explanation of why they deserve their own unique name.
2. Similarly, it is not quite clear
3. While the overall quality of the figures is high, it would be helpful for the reader to have meaningful, descriptive titles added to each panel explaining what is being shown.
4. Line 55: grammar, 'plasmid' should be 'plasmids'
5. Line 131: grammar, 'folds' should be 'fold'
6. At several points throughout the document, the authors refer to Inocle “genomes” (Lines 121, 152 and elsewhere), it is not clear to this reviewer if it is really appropriate to refer to them as genomes, since they are an intracellular extrachromosomal element?
7. Line 189: spelling, should be “nares”, I think
8. Line 190: Would suggest changing “oral site” to “oral cavity”, since the “sites” within the oral cavity are mentioned in the following words.
9. Lines 176-178, please mention the actual number of metagenomic samples from the various countries that were analyzed.
10. Line 326: would suggest changing the wording away from “spreading” to something like “globally distributed”...as “spreading” could perhaps suggest this is only a current/recent distribution across the world, and there doesn't appear to be evidence provided that suggests when these elements appeared and spread.
11. Lines 336-337: would suggest changing to “known antibacterial or virulence genes”, since it would seem quite possible that Inocles do harbor antibacterial and/or virulence genes, however they are unknown, since it is stated later in the discussion that 95% of the genes in Inocles have “elusive” functions.
12. Line 345: grammar, should be “...Inocles are a key factor...”
13. Line 347: grammar, should be “...additional feature of Inocles is that...”
14. It is a bit unclear if there is evidence what specific *Streptococcus* species may be harboring Inocles? Inocle_004 appears to be found in *S. salivarius*, but what about other Inocles? Or if this can only be predicted at a genus level just clearly state that.
15. Figure 3: would be helpful to include the N's of the samples from the various countries right in the figure.

Reviewer #2

(Remarks to the Author)

In the report titled “Giant extrachromosomal element “Inocle” potentially expands the adaptive capacity of the human oral

microbiome" the authors describe the presence and prevalence of "Inocles" as large extrachromosomal elements that code for what seems to be common bacterial physiological functions.

The methodology described to extract the extra chromosomal elements, appears to be straightforward and in line to the goals of the study. Nonetheless, it is worth mentioning that the preNuc treatment, although very applicable, may still introduce bias in the detection of these elements as the correlation between preNuc and no-treatment only approximates 0.83 (Spearman's rho). Related to the processes filtering and assembling reads, it would be valuable to align the contigs to all genbank instead to specific databases (i.e. UniRef90), in order to investigate the potential origin and source of the large extra chromosomal elements. Additionally, it would be beneficial to test alternative assembly methods and compare results with metaFlye.

It has been known that bacteria often encode redundancy in their genomes as a backup system for survival and adaptability. Although the authors employ this rationale to characterize the genetic content of the ECEs, it was not clear if the authors compared it directly to the chromosomal content of the proposed host. Therefore, it is strongly suggested that the authors carry out this comparison to shed light into the origin (e.g. paralogous genes) potential functional redundancy of the ECEs identified.

The authors identify *Streptococcus* as the potential host bacterium based on the presence of one of the "Inocle" taxonomies described via PCR and whole genome shotgun sequencing of a single sample. However, it is somewhat unclear how *Streptococcus salivarius* was determined to be the host as the authors also state that even at a high sequencing depth, no ECE elements were detected in the same colony. The authors go on to explaining that culture conditions used were not conducive to the detection of these ECEs but it is still unclear how the link was made. It is strongly suggested that the authors clarify this point especially when *S. salivarius* is used later to investigate clinical associations.

Lastly, the authors attempt to correlate ECEs with immune and clinical observations. Although correlations are presented at various levels and with different groups, these fail to control for a number of variables that can affect the presence of bacteria and immune markers. It is strongly advised that if the authors would like to consider adding sections that propose correlation with diseases, a more thorough analysis with an appropriate statistical framework is considered. Additionally, these analyses should include orthogonal or complementary validation methodologies.

Reviewer #3

(Remarks to the Author)

The manuscript submitted by Kiguchi and colleagues reports the discovery of a novel family of ECEs named "Inocles" in the human oral microbiome. The authors created the term "Inocle" to stand for "Insertion sequence encoded; oral origin; circle genomic structure. They revealed that the plasmid-like circular genomic elements could be found in 74% of the global population, primarily residing in *Streptococcus* species, particularly *Streptococcus salivarius*. They also explored the diversity of the Inocle groups and suggested their potential functions as disease markers.

Major issues:

1. If I understand correctly, the identification of *Streptococcus salivarius* as the host of Inocles was mostly based on sequence analysis in their own data. To strengthen this finding, I recommend the following:

1) search in *S. salivarius* isolate genomes, especially those of oral origin, to validate that the Inocles are present,
2) experimentally validate the presence of Inocles in *S. salivarius* strains. Specifically, the authors could isolate a few *S. salivarius* strains from their samples, sequence the genomes the isolates and re-identify Inocles from the sequencing results. This would provide robust evidence supporting their claim.

2. While the authors made public their raw data (PRJDB18431) and the InoC gene sequences (both AA and NT, <https://zenodo.org/records/13138926>), the genomic sequences of the Inocles themselves are not found. Although promised in PRJDB18431, only raw sequencing data appear to be available. Without access to the complete genomic sequences of the Inocles, I am unable to fully evaluate their genomic content.

However, I manually selected a few protein sequences of the InoC gene (the first seven to be specific) and performed a BLAST search against the NCBI NR database using blastp. Surprisingly, all seven InoC genes showed best hit with a replication-relaxation family protein from *Parvimonas micra* (accession ID: WP_278476621.1), with some having >99% AA identity covering 100% of the query proteins (456 AAs). Interestingly, the second best hits of all queries in the NR database often had much worse AA identity (<40%). This raises the question which oral species is the host of the Inocles.

While I believe the dataset is valuable and the findings are novel, certain aspects of the analysis cast doubts on the conclusions. Addressing the issues outlined above, particularly through additional experimental validation and improved data accessibility, would greatly strengthen the manuscript's claims.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

In this reviewer's opinion, the authors have thoroughly addressed the concerns raised by the reviewers, and the manuscript has been strengthened considerably.

In the interim of this review process, a high-profile team has released a preprint highlighting that metagenome-assembled genomes from long-read assemblies (such as the ones in this report) suffer from more errors than has perhaps been appreciated by the field (Trigodet, Sachdeva, Banfield, and Eren, "Assemblies of long-read metagenomes suffer from diverse errors"; bioRxiv 2025). The authors of this preprint also have developed some tools that make it fairly easy to check contigs for such assembly errors. Therefore, I would highly suggest that the authors of the present study check the inocele contigs for assembly errors using the tools developed by Trigodet et al. to have a high level of confidence in the inocele contigs being reported. For example, the script 'anvi-script-find-misassemblies' allowed this reviewer to check some of our group's own assemblies and identify previously unnoticed errors within about 15 minutes of work.

Provided the additional data illustrating a lack of spurious assembly in the inocele contigs, this report would be acceptable for publication, in the opinion of this reviewer.

Reviewer #3

(Remarks to the Author)

Most of my previous concerns have been addressed successfully by the authors.

However, I would like to point out that the plasmid pUA140 (https://www.ncbi.nlm.nih.gov/nucore/NC_002757.1) belonging to *Streptococcus mutans*, does not seem to encode any protein sequences homologous to any of the InoC genes reported by the authors.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In this study, the authors report the discovery of Inocles, a novel intracellular plasmid-like extrachromosomal element of *Streptococcus*. These novel features have several interesting properties and appear to have correlations with certain immune cells and proteins, and negative correlations with several cancers. Aside from some sporadic grammatical errors, the report is well-written, the approach is logical and well-conducted, the conclusions are largely supported by the results, and my enthusiasm for this article is high. There are several shortcomings of the current report, and addressing these would strengthen the document. Specific points to address are the following:

We thank the reviewer for suggesting changes to increase the impact and readability of our manuscript. We found the reviewer's comments helpful and believe that incorporating changes in response to these comments has substantially improved the manuscript.

We prepared the point-by-point responses in detail below. In summary, we have:

1. added Figure 4E, which shows the clear difference in the codon usage of Inocles compared with known plasmids, indicating their genetic uniqueness in the bacterial plasmid sequences.
2. edited all grammatical errors you pointed out.
3. added Figure 4C to clearly explain the experimental workflow of isolation of *Streptococcus* strains having Inocle_004.

Reviewer #1-1. It is not entirely clear to this reviewer how Inocles differ from plasmids (other than that recent improvements in long-read sequencing technology enabled their discovery). Clearly Inocles are novel, but it would be helpful for the reader to have a more straightforward explanation of why they deserve their own unique name.

Response:

Thank you for your precious suggestion. We additionally analyzed the codon usage of Inocles, plasmids, chromosomes, and phages of *Streptococcus* and showed that Inocles have a clearly different codon usage from the known plasmids (new **Figure 4E**). In addition, we revised the figure of tetranucleotide frequency to make it easy to figure out the differences in nucleotide composition between the known plasmids and Inocles (revised **Figure 4D**). Based on these results, we have stated that Inocles has a clearly different nucleotide composition and codon usage from known plasmids (Line 228-233) and added a new method description (Line 777-779).

Reviewer #1-2. While the overall quality of the figures is high, it would be helpful for the reader to have meaningful, descriptive titles added to each panel explaining what is being shown.

Response:

In response to your suggestion, we added the headline title in each figure legend (Figure 1B, 1C, 1D, 1E, Figure 2B, 2H, 2I, Figure 3D, Figure 4A, 4C, 4D, 4E, Figure 5I, Figure 6A, and 6B). We show the example of revising the figure legend below.

Before Figure 1C. Total bases of PromethION reads without low-quality and human reads per flow cell acquired from preNuc-treated and -untreated samples.

Revised Figure 1C. Effect of preNuc treatment on the sequence depth of microbial-derived long reads. Bar plots show the total bases of PromethION reads without low-quality and human reads per flow cell acquired from preNuc-treated and -untreated samples.

Reviewer #1-3. Line 55: grammar, 'plasmid' should be 'plasmids'

Response:

Based on the reviewer's comments, we revised the writing (Line 55).

Reviewer #1-4. Line 131: grammar, 'folds' should be 'fold'

Response:

Based on the reviewer's comments, we revised the writing (Line 131).

Reviewer #1-5. At several points throughout the document, the authors refer to Inocle "genomes" (Lines 121, 152 and elsewhere), it is not clear to this reviewer if it is really appropriate to refer to them as genomes, since they are an intracellular extrachromosomal element?

Response:

Thank you for your precise suggestion regarding the definition of "genome." We agree to the careful use of the "genome" for the extrachromosomal element. To avoid confusing the reading, we changed all "genomes" to "sequences" or "contigs." (Line 53, 121, 123, 131, 157, 158, 162, 165, 170, 172, 203, 234, 238, 240, 345, 476, 477, 479, 580, 582, 597, 627, 704, 714, 762, 767, and 784).

Reviewer #1-6. Line 189: spelling, should be "nares", I think

Response:

Based on the reviewer's comments, we revised the writing (Line 190).

Reviewer #1-7. Line 190: Would suggest changing "oral site" to "oral cavity", since the "sites" within the oral cavity are mentioned in the following words.

Response:

Based on the reviewer's comments, we revised the writing (Line 191).

Reviewer #1-8. Lines 176-178, please mention the actual number of metagenomic samples from the various countries that were analyzed.

Response:

Based on the reviewer's comments, we added the number of samples in each country (Line 177-178).

Reviewer #1-9. Line 326: would suggest changing the wording away from "spreading" to something like "globally distributed"...as "spreading" could perhaps suggest this is only a current/recent distribution across the world, and there doesn't appear to be evidence provided that suggests when these elements appeared and spread.

Response:

We agreed to use "globally distributed" instead of "spreading" and revised the writing (Line 335).

Reviewer #1-10. Lines 336-337: would suggest changing to “known antibacterial or virulence genes”, since it would seem quite possible that Inocles do harbor antibacterial and/or virulence genes, however they are unknown, since it is stated later in the discussion that 95% of the genes in Inocles have “elusive” functions.

Response:

Thank you for your precise suggestion related to the antibacterial and virulence genes. Based on the reviewer’s comments, we revised the writing (Line 350).

Reviewer #1-11. Line 345: grammar, should be “...Inocles are a key factor...”

Response:

Based on the reviewer’s comments, we revised the writing (Line 359).

Reviewer #1-12. Line 347: grammar, should be “...additional feature of Inocles is that...”

Response:

Based on the reviewer’s comments, we revised the writing (Line 361).

Reviewer #1-13. It is a bit unclear if there is evidence what specific *Streptococcus* species may be harboring Inocles? Inocle_004 appears to be found in *S. salivarius*, but what about other Inocles? Or if this can only be predicted at a genus level just clearly state that.

Response:

Thank you for your invaluable comments related to the host prediction of Inocles. As the reviewer pointed out, the host bacterial species of Inocles other than Inocle_004 is unclear, so we have concluded that the candidate host bacterial species of Inocles_004 is *Streptococcus salivarius*, but it is still unclear for other Inocles (Line 223-225). Based on this conclusion, we added a sentence in the discussion mentioning our limitation related to predicting host bacterial species in other Inocles (Line 399-402).

Additionally, we performed additional experiments and revised figures and writing to clearly explain the experimental process for identifying the possible host bacterial species of the Inocle_004 (new **Figure 4C**). In this experiment, we investigated the presence of Inocle_004 in five bacterial colonies on an agar plate from the saliva sample (origin of Inocle_004). The sequencing of the PCR product using Inocle_004 specific primers was performed for confirmation of the presence of Inocle_004 in each colony, and we found three colonies have Inocle_004. Interestingly, it was found that Inocle_004 could not be detected after each strain was propagated in liquid culture to obtain sufficient DNA for genomic sequencing. This indicates that Inocle_004 is maintained in the colony but is lost under liquid culture conditions. The 16S rRNA gene sequence (two strains) and whole genome shotgun sequencing (one strain) confirmed that bacterial strains with Inocle_004 (only in colony) were *Streptococcus salivarius*. From these results, we speculate that one of the host bacterial species of Inocle_004 includes *S. salivarius*. We edited and added these isolate-related results to the results (Line 205-227), discussion (Line 398-404), and method sections (Line 678-697).

Reviewer #1-14. Figure 3: would be helpful to include the N’s of the samples from the various countries right in the figure.

Response:

Thank you for your suggestion about Figures. Based on the reviewer’s comments, we added the number of samples in revised **Figure 3D**.

Reviewer #2 (Remarks to the Author):

In the report titled "Giant extrachromosomal element "Inocle" potentially expands the adaptive capacity of the human oral microbiome" the authors describe the presence and prevalence of "Inocles" as large extrachromosomal elements that code for what seems to be common bacterial physiological functions.

We thank the reviewer for suggestions that substantially improved our manuscript, especially the insightful comments regarding the new genomic analysis of host bacteria and Inocles. The comments related to the description of the experimental process also substantially increased the impact of our manuscript. We also appreciated your suggestion related to the statistical framework. We believe the new experiment and analysis increased the confidence in host prediction and in the associations between Inocles and human clinical significance.

We prepared the point-by-point responses in detail below. In summary, we have:

1. added Figure 4C to clearly explain the experimental workflow of isolation of *Streptococcus* strains, including Inocle_004.
2. added Figure S9 and Figure S1 to show the homologous genes between Inocles and candidate host bacterial genomes, suggesting the biological contribution of Inocles for functional redundancy of stress tolerance of *Streptococcus*.
3. reanalyzed the correlation between Inocles and human blood cell population and proteome considering the confounders, and revised Figures 6A and 6B.

Reviewer #2-1. The methodology described to extract the extra chromosomal elements, appears to be straightforward and in line to the goals of the study. Nonetheless, it is worth mentioning that the preNuc treatment, although very applicable, may still introduce bias in the detection of these elements as the correlation between preNuc and no-treatment only approximates 0.83 (Spearman's rho).

Response:

Thank you for your suggestions regarding the technical description. Based on reviewer's comment, we have added to the discussion related to the bias of preNuc treatment for the composition of extrachromosomal elements (Line 339-344).

Reviewer #2-2. Related to the processes filtering and assembling reads, it would be valuable to align the contigs to all genbank instead to specific databases (i.e. UniRef90), in order to investigate the potential origin and source of the large extra chromosomal elements.

Response:

Thank you for your invaluable suggestion for identifying the potential origin of Inocles. We mapped the short-read sequences of >3,000 oral isolates deposited in NCBI to Inocle contigs in order to find the possible origin of Inocles, but we could not find the sample in which Inocle was detected. We have revised the results and method sections to describe these results (Line 205-208 and 823-825).

From the additional experiment to isolate salivary bacteria having Inocle, we speculated the possible reason why genomic sequences from isolated bacteria did not have the Inocle-derived sequences (see response to #2-5).

Reviewer #2-3. Additionally, it would be beneficial to test alternative assembly methods and compare results with metaFlye.

Response:

Thank you for your useful suggestions regarding the assembly aspects. We used metaFlye assembler, which is known to be the most efficient long-read metagenomic assembler for constructing plasmids (Buttler et al., 2022, *Microorganisms*). Considering that Inocle is a plasmid-like element, we think that the use of metaFlye is likely to be an appropriate choice.

<Reference>

Buttler, Jeremy, and Devin M. Drown. 2022. "Accuracy and Completeness of Long Read Metagenomic Assemblies." *Microorganisms* 11 (1): 96.

Reviewer #2-4. It has been known that bacteria often encode redundancy in their genomes as a backup system for survival and adaptability. Although the authors employ this rationale to characterize the genetic content of the ECEs, it was not clear if the authors compared it directly to the chromosomal content of the proposed host. Therefore, it is strongly suggested that the authors carry out this comparison to shed light into the origin (e.g. paralogous genes) potential functional redundancy of the ECEs identified.

Response:

Thank you for your insightful suggestion regarding the relationship between Inocle and host bacterial chromosomes. According to the reviewer's suggestion, we searched for homologous genes between Inocles and *Streptococcus* chromosomes. On average, 20 genes (6% on average)/Inocle showed significant homology (e-value<1e-10) with genes in *Streptococcus* chromosomes (new **Figure S9**). Functionally annotatable genes of Inocles, such as ATP-dependent RecD-like DNA helicase, bifunctional glycosyltransferase, and sortase A, are included in these homologous genes (new **Figure 5I**). These results suggest that the majority of functions of Inocles are unique, but some parts contribute to the functional redundancy of stress tolerance-related systems. According to these results, we added new results (**Figure S9 and Figure 5I**), discussion, and method in the revised manuscript (Line 274-279, 373-378, and 803-805).

Reviewer #2-5. The authors identify *Streptococcus* as the potential host bacterium based on the presence of one of the "Inocle" taxonomies described via PCR and whole genome shotgun sequencing of a single sample. However, it is somewhat unclear how *Streptococcus salivarius* was determined to be the host as the authors also state that even at a high sequencing depth, no ECE elements were detected in the same colony. The authors go on to explaining that culture conditions used were not conducive to the detection of these ECEs but it is still unclear how the link was made. It is strongly suggested that the authors clarify this point especially when *S. salivarius* is used later to investigate clinical associations.

Response:

Thank you for your important comments relating isolation of bacteria with Inocle. We have added a figure to describe the culture experiment more clearly and have revised the writing (new **Figure 4C**). Through our additional experiments, we investigated the presence of Inocle_004 in five bacterial colonies on an agar plate directly cultured from the saliva sample of the origin of Inocle_004. The sequencing of the PCR product using Inocle_004 specific primers was performed for confirmation of the presence of Inocle_004 in each colony, and we found three colonies have Inocle_004. Notably, it was found that Inocle_004 could not be detected after each strain was propagated in liquid culture to obtain sufficient DNA for genomic sequencing. This indicates that Inocle_004 is maintained in the colony but is lost under liquid culture conditions. We interpret this to be why we could not find Inocle-derived sequences from our genome sequencing data and the sequences deposited in NCBI (related to the response of #2-2). The whole genome shotgun sequencing (one strain) and 16S rRNA gene sequencing (two strains) confirmed that bacterial strains with Inocle_004 (only in colony) were *Streptococcus salivarius*. From these results, we speculate that one of the host bacterial species of Inocle_004 includes *S. salivarius*. The host bacterial species of Inocles other than Inocle_004 are unclear, so we have revised the conclusion of the candidate host of Inocles as *Streptococcus* and the possible host bacterial species of Inocle_004 is *S. salivarius*. We have revised the result (Line 208-227), discussion

(Line 398-404), and method sections (Line 678-697). Based on these conclusions, we conducted a correlation analysis between the *Streptococcus* and human physiological datasets in a later part to investigate clinical associations (revised **Figure 6**).

Reviewer #2-6. Lastly, the authors attempt to correlate ECEs with immune and clinical observations. Although correlations are presented at various levels and with different groups, these fail to control for a number of variables that can affect the presence of bacteria and immune markers. It is strongly advised that if the authors would like to consider adding sections that propose correlation with diseases, a more thorough analysis with an appropriate statistical framework is considered. Additionally, these analyses should include orthogonal or complementary validation methodologies.

Response:

Thank you for your insightful comments on the confounders of correlation analysis. In response to the reviewer's comments, we considered the statistical framework of the correlations between 1) Inocles and immunity and 2) Inocles and clinical observations accounting for demographic variables that are known to affect human commensal composition and immune status. We reflected the results in the revised manuscript (revised **Figure 6**) and added a new method description (Line 856-861). Specifically, the following corrections have been made.

1) Inocles and immunity

1-1) Correlation analysis between Inocle- α and PBMC cell types.

The partial Spearman's correlation has been applied to identify significant correlations between the human microbiome and host physiological variables, such as plasma metabolites, accounting for confounders (Dekkers et al., 2022, *Nature Communications*; Takeuchi et al., 2023, *Nature*). Therefore, we reanalyzed the correlation between Inocle- α and PBMC cell types (19 cell types) using a partial Spearman's correlation method that accounts for age and sex as confounders. As a result, we confirmed significant positive correlations between three B cell types and Inocle- α in the new analysis (revised **Figure 6A**). Also, we confirmed a significant negative correlation between CD16+ and CD14+ Monocyte and Inocle- α . Based on these results, we revised the results section (Line 290-295). The original Figure 6B has been deleted in the revised manuscript to avoid a statistical test that does not consider the confounders.

1-2) Correlation analysis between Inocle- α and plasma proteome.

We also reanalyzed the correlation between the plasma proteome dataset (Olink technology) and Inocle- α using partial Spearman's correlation method, accounting for age and sex as confounders. It was confirmed that infection-related pathways and cancer-related pathways were positively correlated with Inocle- α (Line 298-309, revised **Figure 6B and Table S11**). Importantly, the B cell receptor signaling pathway in plasma proteome was positively correlated with Inocle- α (revised **Figure 6B**). We speculate that this result corresponds to the positive correlation between B cells and Inocle- α (revised **Figure 6A**). We believe that these results emphasize the relationship between B cells and Inocle- α through the application of two different analytical methods (Line 302-303).

1-3) Correlation analysis between Inocle- α and gene expression in B cells of PBMC.

In terms of correlation analysis between gene expression of B cells and Inocle- α , we also conducted partial Spearman's correlation analysis, accounting for age and sex as confounders. However, we could not observe any significantly correlated genes. These results suggest that gene expression in B cells is influenced by confounders. To avoid results not accounting for confounders, the original Figure 6C and Table S9 have been deleted.

2) Inocles and clinical observations

In the statistical analysis to clarify the association between HNC patients and Inocle- α , confounders (age, sex, smoking status, and pre/post-cancer treatment) were considered using Maaslin2, which is the standard case-control comparison framework in microbiome science. In terms of CRC, PDAC, and RA, we considered all publicly available confounders, such as age and sex, using Maaslin2. This indicates that the effects of the available confounders have been minimized. However, as the reviewer pointed out, we could not consider the many suspected other confounders (e.g., diet and lifestyle) in the human microbiome and immune status, so we have included a discussion of the limitations controlling confounders (Line 391-394). To enhance the clarity of the figure's interpretation, we have revised the correlation analysis figure that compares the abundance of Inocle or *Streptococcus* in the saliva of healthy individuals and those with the disease (revised **Figure 6E**). We have also added information on the confounders to the figure legend for easy understanding of the considered confounders (Line 541-542 and 543-545).

We agree that it is important to validate our observation by complementary methods. The correlation analysis using other independent validation cohorts is the most robust way for validating the association between the microbiome and disease and has been applied to several studies (Wirbel et al. 2019. *Nature Medicine*, Nagata et al. 2022. *Gastroenterology*). However, this study is based on our original dataset, and there are no previous studies that generated similar data sets of salivary shotgun metagenome, blood scRNA-seq, and plasma proteome. We found the previous study of salivary shotgun metagenome from head and neck cancer patients (Wu et al. 2023. *Oral Oncology*), but shotgun metagenomic sequences were not publicly available (only ITS amplicon data is available), which made it difficult to validate our findings using other cohorts. Therefore, we added a discussion to mention the importance of follow-up studies for validating the associations between Inocles and clinical observations (Line 394-396).

<Reference>

Dekkers, Koen F., Sergi Sayols-Baixeras, Gabriel Baldanzi, Christoph Nowak, Ulf Hammar, Diem Nguyen, Georgios Varotsis, et al. 2022. "An Online Atlas of Human Plasma Metabolite Signatures of Gut Microbiome Composition." *Nature Communications* 13 (1): 5370.

Takeuchi, Tadashi, Tetsuya Kubota, Yumiko Nakanishi, Hiroshi Tsugawa, Wataru Suda, Andrew Tae-Jun Kwon, Junshi Yazaki, et al. 2023. "Gut Microbial Carbohydrate Metabolism Contributes to Insulin Resistance." *Nature* 621 (7978): 389–95.

Wirbel, Jakob, Paul Theodor Pyl, Ece Kartal, Konrad Zych, Alireza Kashani, Alessio Milanese, Jonas S. Fleck, et al. 2019. "Meta-Analysis of Fecal Metagenomes Reveals Global Microbial Signatures That Are Specific for Colorectal Cancer." *Nature Medicine* 25 (4): 679–89.

Nagata, Naoyoshi, Suguru Nishijima, Yasushi Kojima, Yuya Hisada, Koh Imbe, Tohru Miyoshi-Akiyama, Wataru Suda, et al. 2022. "Metagenomic Identification of Microbial Signatures Predicting Pancreatic Cancer From a Multinational Study." *Gastroenterology* 163 (1): 222–38.

Wu, Zeni, Yongli Han, Yunhu Wan, Xing Hua, Samantha S. Chill, Kedest Teshome, Weiyin Zhou, et al. 2023. "Oral Microbiome and Risk of Incident Head and Neck Cancer: A Nested Case-Control Study." *Oral Oncology* 137 (106305): 106305.

Reviewer #3 (Remarks to the Author):

The manuscript submitted by Kiguchi and colleagues reports the discovery of a novel family of ECEs named "Inocles" in the human oral microbiome. The authors created the term "Inocle" to stand for "Insertion sequence encoded; oral origin; circle genomic structure. They revealed that the plasmid-like circular genomic elements could be found in 74% of the global population, primarily residing in *Streptococcus* species, particularly *Streptococcus salivarius*. They also explored the diversity of the Inocle groups and suggested their potential functions as disease markers.

We thank the reviewer for the comments that improved our manuscript, especially regarding the host prediction of Inocles. We also appreciated the comments related to sharing our dataset. We believe that incorporating changes in response to these comments has substantially improved the manuscript.

We prepared the point-by-point responses in detail below. In summary, we have:

1. performed additional genomic analysis to identify the Inocle-derived sequences from orally isolated bacterial strains in the NCBI database.
2. performed additional experiments to identify the *Streptococcus* strains having Inocle and added Figure 4C to explain the experimental workflow of isolation.
3. checked data sharing status and confirmed that all datasets are publicly available.

Major issues:

If I understand correctly, the identification of *Streptococcus salivarius* as the host of Inocles was mostly based on sequence analysis in their own data. To strengthen this finding, I recommend the following:

Reviewer #3-1. search in *S. salivarius* isolate genomes, especially those of oral origin, to validate that the Inocles are present,

Response:

Thank you for your invaluable suggestion for identifying the potential origin of Inocles. We mapped the short-read sequences of >3,000 oral isolates deposited in NCBI to Inocles in order to find the possible origin of Inocles, but we could not find the sample in which Inocle was detected. We have revised the result and method sections to describe these findings (Line 205-208 and 823-825).

From the additional experiment to isolate salivary bacteria having Inocle, we speculated the possible reason why genomic sequences from isolated bacteria did not have the Inocle-derived sequences (see below response).

Reviewer #3-2. experimentally validate the presence of Inocles in *S. salivarius* strains. Specifically, the authors could isolate a few *S. salivarius* strains from their samples, sequence the genomes the isolates and re-identify Inocles from the sequencing results. This would provide robust evidence supporting their claim.

Response:

Thank you for your invaluable suggestion to confirm the host bacterial species of the Inocle. We have added a figure (new **Figure 4C**) to describe the culture experiment more clearly. Through our additional experiments, we investigated the presence of Inocle_004 in five bacterial colonies on an agar plate directly cultured from the saliva sample of the origin of Inocle_004. The sequencing of the PCR product using Inocle_004 specific primers was performed for confirmation of the presence of Inocle_004 in each colony, and we found three colonies have Inocle_004. Notably, it was found that Inocle_004 could not be detected after each strain was propagated in liquid culture to obtain sufficient DNA for genome sequencing. This

indicates that Inocle_004 is maintained in the colony but is lost under liquid culture conditions. We interpret this to be why we could not find Inocle-derived sequences from our genome sequencing data and the sequences deposited in NCBI. The whole genome sequencing (one strain) and 16S rRNA gene sequencing (two strains) confirmed that bacterial strains with Inocle_004 are *Streptococcus salivarius*. From these results, we conclude that one of the host bacterial species of Inocle_004 includes *S. salivarius*, but the host bacterial species of Inocles other than Inocle_004 is unclear (Line 223-225). Based on these experimental results, we have revised the result (Line 208-227), discussion (Line 398-404), and method sections (Line 678-697).

Reviewer #3-3. While the authors made public their raw data (PRJDB18431) and the InoC gene sequences (both AA and NT, <https://zenodo.org/records/13138926>), the genomic sequences of the Inocles themselves are not found. Although promised in PRJDB18431, only raw sequencing data appear to be available. Without access to the complete genomic sequences of the Inocles, I am unable to fully evaluate their genomic content.

Response:

Thank you for your valuable comments about the data sharing. We have deposited all our data on DDBJ, but it seems like it has not been synchronized with NCBI yet. We have attached the URL where the data can be accessed below. We added this link to the method section, too (Line 631).

DDBJ URL: <https://ddbj.nig.ac.jp/search/entry/sra-run/DRR582113>

Inocle contig file: DRZ139060, DRZ139061, DRZ139062, DRZ139063, DRZ139064, DRZ139065, DRZ139066, DRZ139067, DRZ139068, DRZ139069, DRZ139070, DRZ139071, DRZ139072, DRZ139073, DRZ139074, DRZ139075, DRZ139076, DRZ139077, DRZ139078, DRZ139079, DRZ139080, DRZ139081, DRZ139082, DRZ139083, DRZ139084, DRZ139085, DRZ139086, DRZ139087, DRZ139088

Reviewer #3-4. However, I manually selected a few protein sequences of the InoC gene (the first seven to be specific) and performed a BLAST search against the NCBI NR database using blastp. Surprisingly, all seven InoC genes showed best hit with a replication-relaxation family protein from *Parvimonas micra* (accession ID: WP_278476621.1), with some having >99% AA identity covering 100% of the query proteins (456 AAs). Interestingly, the second best hits of all queries in the NR database often had much worse AA identity (<40%). This raises the question which oral species is the host of the Inocles.

Response:

Thank you for your interest in our dataset and insightful feedback regarding host identification. We investigated the reference publication linked to the WP_278476621.1 and found that the study sequenced a plasmid (pUA140) from *Streptococcus mutans* (Zou et al., 2001, *Plasmid* and Link1 in reference below). But this gene is also linked to the metagenome-assembled genome (MAG) of *Parvimonas micra* (Link1 and Link2). It is unclear why WP_278476621.1 is registered as a protein of *P. micra* MAG, although the paper is cited on the plasmid of *S. mutans*. In the revised manuscript, we have presented more specific data to support the conclusion that *S. salivarius* is a candidate host bacterial species for Inocle_004 (new **Figure 4C**). Furthermore, we revised to clearly state that the host species for Inocles except Inocle_004 are currently unknown (Line 224–225), and our conclusions do not exclude hosts outside the *Streptococcus* genus (Line 402–403). We expect that further experimental studies will identify a more definitive host range for Inocles (Line 403-404).

<Reference>

Zou, X., P. W. Caufield, Y. Li, and F. Qi. 2001. "Complete Nucleotide Sequence and Characterization of pUA140, a Cryptic Plasmid from *Streptococcus Mutans*." *Plasmid* 46 (2): 77–85.

Link1: https://www.ncbi.nlm.nih.gov/protein/WP_278476621.1/

Link2: https://www.ncbi.nlm.nih.gov/ipg/?term=WP_278476621

Link3: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_015263295.1/

Reviewer #1 (Remarks to the Author)

In this reviewer's opinion, the authors have thoroughly addressed the concerns raised by the reviewers, and the manuscript has been strengthened considerably.

In the interim of this review process, a high-profile team has released a preprint highlighting that metagenome-assembled genomes from long-read assemblies (such as the ones in this report) suffer from more errors than has perhaps been appreciated by the field (Trigodet, Sachdeva, Banfield, and Eren, "Assemblies of long-read metagenomes suffer from diverse errors"; bioRxiv 2025). The authors of this preprint also have developed some tools that make it fairly easy to check contigs for such assembly errors. Therefore, I would highly suggest that the authors of the present study check the inoceles for assembly errors using the tools developed by Trigodet et al. to have a high level of confidence in the inocele contigs being reported. For example, the script 'anvi-script-find-misassemblies' allowed this reviewer to check some of our group's own assemblies and identify previously unnoticed errors within about 15 minutes of work.

Provided the additional data illustrating a lack of spurious assembly in the inocele contigs, this report would be acceptable for publication, in the opinion of this reviewer.

Response:

Thank you for your invaluable comment on the quality of the Inocele contigs. To check for possible misassembly of Inocele contigs, we mapped long reads to individual Inocele contigs using minimap2. Then, we ran the 'anvi-script-find-misassemblies' to identify the zero coverage regions and 100% clipping regions (--clipping-ratio 1.0), which are mentioned as the confident errors of long-read assembly (Trigodet, et al., bioRxiv 2025). We confirmed that none of the contigs contained unsupported single-nucleotide variants, insertions, deletions, or clipped regions. We believe this result is because we carefully confirmed that there were no strange mapping results in the Inocele contigs, as shown in Figure S3. These results indicate that there were no assembly errors in any of the Inocele contigs. We have deposited all datasets below.
DOI 10.5281/zenodo.15557843 (<https://zenodo.org/records/15557844>)

Reviewer #3 (Remarks to the Author)

Most of my previous concerns have been addressed successfully by the authors.

However, I would like to point out that the plasmid pUA140 (https://www.ncbi.nlm.nih.gov/nuccore/NC_002757.1) belonging to *Streptococcus mutans*, does not seem to encode any protein sequences homologous to any of the InoC genes reported by the authors.

Response:

We appreciate the reviewer for multifaceted and invaluable comments on the significant data for host prediction of Inocles. We re-examined the *P. micra* genome encoding InoC carefully evaluated the similarity between InoC and WP_278476621.1. This genome is a metagenome-assembled genome constructed using saliva metagenome sequencing, and the 1,371 bp contig (NZ_JABZRE010000001.1) encoding InoC has been classified into the same bin as *P. micra* by binning analysis. This contig was highly similar to Inocele_022 (>99% identity), indicating that this contig was derived from Inocele. However, to our knowledge, there is no evidence that binning analysis accurately predicts the host bacterium of the plasmid, and a review article assessing metagenomic technology has pointed out that coincidentally, plasmid or phage contigs are classified into the chromosomal bin as a principal limitation of binning analysis¹. Based on this knowledge, it is unclear whether *P. micra* is included as the host of Inocele.

We believe that our interpretation regarding the host bacterium of Inocles was improved through the review process, and we agree with the reviewer's opinion that the host of Inocles cannot yet be definitively determined. We reviewed our manuscript again and revised it to avoid stating the specific host bacteria of

Inocles. Specifically, we have revised the writing in the summary section to avoid stating that the host of Inocles is *Streptococcus*, but instead, we state that it includes *Streptococcus* (Line: 39-40).

Reference

Chen, Lin-Xing, Karthik Anantharaman, Alon Shaiber, A. Murat Eren, and Jillian F. Banfield. 2020. "Accurate and Complete Genomes from Metagenomes." *Genome Research* 30 (3): 315–33.

Comments on behalf of Reviewer #2:

In my opinion, the authors have sufficiently addressed the concerns raised by Reviewer #2. For points 2-5 and 2-6, specifically, they provide extensive commentary in their Response Letter and have modified the Main Text document accordingly.

Response:

We appreciate your comments on behalf of Reviewer #2 and are glad to know that our effort have addressed the concerns raised by Reviewer #2.